통 계 조 사 실 습

#1 자유응답형 문항을 만들 때 고려사항은?

자유응답형 질문은 응답자가 응답을 하도록 준거틀을 제공하지만, 응답의 표현 방식이나 내용을 제한하지 않는 방식이다. 깊이 파고 들어갈 수 있으며, 응답자의 모호한 반응을 탐지하고 라포를 형성하게 해주고, **응답자의 진정한 의도, 신념 및 태도를 보다 잘 평가**해준다. 그러나 **다양한 응답을 가져옴으로써 자료 처리를 위한 부호화가 어렵다.** 또한 응답자에게 지나친 부담을 안겨줄 수 있다.

#2 고정선택형 문항을 만들 때 고려사항은?

고정선택형 문항은 집계 및 분석 시 매우 편리하며, 응답자들에게 가능하나 응답을 알려줌으로써 응답 상 오류를 최소화한다. 그러나 다지선다형 문제를 만들기 위한 **비용과 시간이 많이 들며, 응답 지문의 제시 순서가 결과의 영향을 미친다.** 또한 고전선택형 문항을 만들기 위해서는 **응답 지문들은 상호배타적이고, 응답자의 가능한 모든 응답을 포괄하도록** 작성해야 한다.

#3 설문지 작성 시 주의사항은?

질문 작성 시 보편적이고 상용적인 용어를 사용하여 정확하게 질문을 파악할 수 있게 한다. 질문은 객관적이고 긍정/부정 어느 한 방향으로 치우쳐선 안 되며, 질문 내의 어떤 가정이나 암시가 있었어도 안 된다. 질문은 가능한 간단해야 하며, 질문의 용어는 가치중립적이어야 하며, 위험한 용어, 인기 용어 등은 피해야 한다. 신망 있는 사람이나 악명 높은 사람의 이름을 인용할 경우에는 신중을 가해야 하며 응답자가 알리기를 꺼려하는 신상의 문제나, 지위, 현황 등을 묻는 질문은 어구 구성에 있어 응답자가 심리적으로 안심할 수 있도록 구성해야 한다. 질문에서 찬성과 반대의 뜻으로 대답하는 카테고리 수는 같아야 하며, 질문이 애매하거나 막연한 내용이 포함되지 않아야 한다. **질문의 순서는 첫 질문은 간단하고, 흥미 있는 질문, 응답자가 쉽게 대답할 수 있는 질문으로 한다.** 일반적인 질문 후에 특수한 질문을 뒤에 배열한다. 오래된 것부터 최근의 것으로, 또는 최근의 것부터 오래된 것으로 배열한다. 앞에 있는 질문이 뒤에 올 질문의 대답에 영향을 줄 수 있는 경우는 되도록 피하도록 한다.

#4 리커트형 척도와 5점 척도의 공통점과 차이점은?

5점 척도는 등급척도로 어떤 측정 대상이 가지고 있는 속성의 정도를 그림, 글 혹은 숫자로 평가하는 방법이다. 가장 흔히 사용되는 척도로 만들기 쉽고 간편하며 시간과 비용이 매우 경제적이다. 그러나 평가자의 성격이나 태도 등의 성향에 따라 편향된 결과가 나올 수 있다. 리커트형 척도는 응답자들에게 각 문항에 대한 동의 또는 반대의 척도를 표현하도록 측정하는 방법이다. 등간척도보다는 서열 척도에 가깝지만 서열척도 중 등간격을 나타 낼 수 있는 가장 좋은 척도이다. **5점 척도와 리커트형 척도는 외관상으로는 비슷해 보이지만 5점 척도는 등간격이 아니지만 리커트형 척도는 등간격으로 간주하는 것이 가장 큰 차이**이다.

#5 여러 가지 척도에 대하여 설명하여라.

서열척도를 이용한 척도에는 **순위법**, **범주별순위할당법**이 있다. 순위법은 어떤 속성에 대한 순위를 기록하여 측정하는 방법이고, 범주별순위할당법은 비교대상이 많을 때 이용가능하다. 척도법을 이용한 측정 방법에는 **등급 척도**, **리커트형 척도**, **유사등간척도**, **어의차별척도**, **스타펠 척도** 등이 있다. 유사등간척도는 태도 측정에 사용되는 평위척도로 적합-중립-부적합의 세분화로 주로 11점 척도로 구성한다. 어의차별척도는 응답자에게 극단적인 형용사를 제시하여 응답자의 느낌과 가장 가까운 부분을 표시하도록 한다. 스타펠 척도는 어의차별척도의 변형으로 평가 기준으로 양극단 대신 중간에 하나의 수식어만 제시한다. 비율척도를 이용한 측정에는 **총합고정척도**와 **비율분할법**이 있다. 총합고정척도법은 응답자에게 고정된 총점수를 제시하여 속성들의 상대적 중요도를 점수로 평가하는 측정이고, 비율분할법은 응답자들에게 속성들의 수준을 기준으로 다른 속성과 상대적으로 비교해 측정하는 방법이다.

#6 데이터의 수준에 대하여 설명하시오.

명목척도는 대상의 상황, 상태, 분류를 측정하는 척도로 질적자료(비수치)이다. **서열척도**는 대상들에 대해 순서, 개념을 이용한 측정이며 질적자료(일부 수치)이다. **등간척도**는 속성 차이 비교가 가능하도록 수치로 측정하며 상대 0점이 존재한다. **비율척도**는 속성 차이 및 비율 비교가 가능한 수치로 절대/상대 0점이 존재한다.

명목척도 : =, ≠
순서척도 : =, ≠, 〉, 〈
등간척도 : =, ≠, 〉, 〈, +, −
비율척도 : =, ≠, 〉, 〈, +, −, ×, ÷

#7 보고서 작성 시 주의사항은?

보고서 작성 시 **보고 대상을 명확히 파악**해야 하며, **보고내용의 핵심을 명확**하게 해야 한다. **보고목적은 확실**해야 하며, **보고시기를 엄수**해야 한다. 또한 보고서 작성 시 이용자의 관점에서 보고서 이용자가 쉽게 이해할 수 있게 해야 하며, 작성자의 이해관계나 주관이 배제된 객관성이 유지되어야 한다. 보고서 그 자체로 내용과 형식면에서 완성도가 있어야 한다. 양식은 간결하고 명료하고 효율적인 양식을 갖추어야 한다. 또한 통계 보고서 작성 시에는 **간단명료한 표현**을 사용하며, **능동태와 수동태를 구분**하여 쓰며, **생략과 약자는 삼가도록** 한다. **일관된 용어를 사용**해야 하며 비록 나머지 범주가 넓고 가치가 없다 하더라도 '기타'라는 **애매모호한 용어는 최대한 피하도록** 한다. '지난 해' 등과 같은 **애매모호한 날짜 표현도 삼간다.** 자료가 인용되거나 언급될 때마다 **기준시기를 명확히 표기**해주며, **일관성 있고 단순화된 비율**을 사용하도록 한다.

#8 가중치에 대하여 설명하시오.

가중치는 표본자료분석에서 **불균등선택확률을 보정**하고, **무응답을 보정**하며, **사후층화를 위해** 사용된다. 일반적으로 최종가중치는 불균등선택확률의 가중치(w_1), 표본의 무응답(w_2), 사후층화를 위한 가중치(w_3)를 곱한 $w_F = w_1 \times w_2 \times w_3$ 로 계산한다.

$$W = \frac{1}{\text{추출율}} \times \frac{\text{목표표본 수}}{\text{응답자 수}} \times \frac{\text{중요집단의 모집단 수}}{\text{중요집단의 가중치}}$$

만약 복합표본조사를 단순확률표본으로 분석하면 단순확률표본 가정에 위배되며, 표준의 통계패키지를 이용한 분석이 어렵고 가중치 미반영시 편향된 결과를 산출하며, 분산이 과소추정된다. **가중치를 부여하면 분산은 커지지만 편향은 제거**된다.

#9 빈도분석에서 독립성 검정과 동일성 검정에 대하여 설명하시오.

독립성 검정은 두 가지 속성(변수)이 서로 독립인지 아닌지를 검정하는 것이다. **귀무가설은 "두 변수는 독립이다."** 이며 대립가설은 "두 변수는 독립이 아니다." 이다. 이 때 검정통계량은 χ^2 검정통계량을 사용하며 $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$ (단, o_{ij}는 관측값, e_{ij}는 기대값이며, c는 행의 수, r은 열의 수) 이다.

동일성 검정은 두 개 또는 두 개 이상의 모집단이 어떤 특성을 갖는 분포에 대하여 서로 동질인가를 검정하는 것이다. **귀무가설은 "각 집단은 동일한 분포를 따른다."** 이고 대립가설은 "각 집단은 동일한 분포를 따르지 않는다." 이다. 이 때 검정통계량은 χ^2 검정통계량을 사용하며 $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$ (단, o_{ij}는 관측값, e_{ij}는 기대값이며, c는 행의 수, r은 열의 수) 이다. 독립성 검정과 동일성 검정의 검정통계량 값은 차이가 없지만 가설과 해석에서의 차이가 있다.

#10 집단분석에서 사후분석을 하는 이유는 무엇인가?

사후분석, 사후검정, 다중비교은 집단 분석에서 **집단 간의 차이가 유의하게 나왔을 때, 집단을 분리시켜 각각의 차이의 정도를 보고자 할 때** 사용한다. t-검정과 분산분석에서의 귀무가설은 "집단 간의 평균 차이가 없다." 이다. 따라서 각 집단 별로 얼마가 차이가 나는지는 알 수 없다. 그렇기 때문에 어떤 집단이 이질적인 집단인지 차이가 얼마나 나는지 Duncan, Tukey, Bonferroni, SNK 검정 등을 통해 집단 간의 차이의 정도를 비교한다.

#11 층화추출 데이터를 회귀 분석할 때 회귀계수의 추정방법에 대하여 설명하시오.

층화추출된 데이터에서 회귀분석을 할 때 회귀 계수를 추정하는 방법은 2가지이다. 첫 번째는 분리회귀계수 추정법이고, 두 번째는 결합회귀계수 추정법이다. 분리회귀계수 추정법은 층간의 차이가 명확할 때 사용 하는 방법으로 $\hat{\beta}_S = \frac{N_M}{N}\hat{\beta}_M + \frac{N_F}{N}\hat{\beta}_F$ 등의 방법으로 회귀 계수를 추정한다. 결합회귀계수 추정법은 층간의 차이가 크지 않을 때 사용되는 방법으로 SAS, SPSS 등은 이 방법으로 회귀계수를 추정한다. $\hat{\beta} = \frac{S_{XY}}{S_{XX}} = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2}$

#12 중복응답을 데이터 입력하는 방법에 대하여 설명하시오.

중복응답은 하나의 질문에서 두 개 이상의 응답을 유도하는 문항이다. 중복 응답 유형에는 순위 응답 질문 유형, 중복 응답 질문 유형, 다중 이분형 응답 질문 유형이 있다. 이런 중복응답을 분석할 때는 **만약 항목이 8개라면 응답1, 응답2, ..., 응답8 의 8개의 변수를 생성하고 각 항목 별로 '0'과 '1'의 이분형으로 코딩**을 한다. 각 응답별로 체크된 응답은 '1' 체크가 안 된 응답은 '0' 으로 표기한다.

#13 무응답과 결측값에 대하여 설명하시오.

무응답은 마땅히 응답을 해야 하는데 응답하지 않은 경우를 나타내며 **결측값은 응답할 필요가 없기 때문에 응답하지 않았음**을 나타낸다. 무응답은 단위무응답과 항복무응답으로 나누며 단위무응답은 설문 자체를 응시하지 않은 경우로 표본을 대체하거나 가중치를 두어 보정한다. 항목 무응답은 9,99,999 등으로 표기하여 그대로 두거나, 평균대체, 회귀대체 등의 방법으로 대체하기도 한다.

#14 데이터의 신뢰도와 일관성을 설명하시오.

신뢰성 문항 검토란 응답자가 문항에 거짓 없이 응답하는지 확인하기 위해 작성된 문항으로 설문지 내 서로 유사한 문항을 작성해 확인하는 문항이다. 유 사 문항에 대한 응답이 다른 경우, 응답자가 설문 문항에 전반에 대해 거짓 혹은 편향되게 응답한 것으로 판단해 분석 자료에서 제외하는 것이 바람직하다. **일치성 문항 검토 및 수정은 응답자가 수치 자료로 응답하는 경우, 수치 자료의 일치성을 검토해 수정하는 과정**이다. 일반적으로 응답자의 응답 자료를 기준으로 판단하는 것이 바람직하다.

#15 데이터 차원을 축소하는 방법에 대해 설명하시오.

데이터 차원을 축소하는 기법에는 변수를 축약하는 기법이 있고, 자료를 축약하는 기법이 있다. 변수를 축약하는 기법에는 주성분분석과 요인분석이 있고, 자료를 축약하는 기법에는 군집분석이 있다. **주성분분석과 요인분석은 어떤 개념을 직접ㅂ측정하기 어려울 때, 그 개념이 공통적으로 내포된 여러 개의 개별 문항을 측정하여 공통 속성이 내포된 변수로 축약하는 과정**으로 비 측정된 새로운 가공 변수를 생성할 때 주로 사용하는 방법이다. **상관계수가 높은 변수끼리 Grouping 하는 방법은 요인분석**이고, **각 변수가 독립이 되도록 변수를 축약하는 방법은 주성분분석**이다. **군집 분석은 관심을 가지고 있는 변수들을 기준으로 특성이 유사한 응답(조사) 대상을 하나의 집단으로 만들어 줌**으로써 조사대상에 대해 그룹을 생성하는 방법이다.

#16 통계 품질은 어떤 속성으로 평가하는가?

품질이란 제품의 규격이 얼마나 잘 들어맞는가를 의미한다. 따라서 **통계에서 품질이란 얼마나 오차가 적은가를 평가**하는 것이다. 그러므로 추정하고자 하는 모집단의 특성과 추정된 특성값과의 차이가 작을수록, 즉 오차가 적을수록 통계의 품질이 좋다고 말할 수 있다. 그렇다면 오차를 어떻게 줄일 수 있을까? 표본조사에서는 표본 차와 비표본오차가 존재하는데 **표본오차는 표본이 모집단에 대한 완전한 정보를 가지고 있지 않으므로 표본을 통해 모수를 추정할 때 발생되는 오차** 이고, 이 오차를 줄이기 위해 표본설계를 통해 어느 정도 통제가 가능하다. 비표본오차는 실제 조사 과정이나 무응답, 설문지의 표현, 자료의 입력, 분석과 같은 자료의 처리 과정에서 발생되는 오차로써 표본오차 이외의 모든 오차를 의미하며 기본적으로는 통제가 불가능한 오차이다. 전수조사에서는 표본오차는 존재하지 않지만 비표본오차는 존재하며, 표본조사에서는 표본오차와 비표본오차 모두 존재한다.

#17 한신대학교 4학년을 대상으로 희망 취업 분야에 대해 조사하고자 한다. 조사계획을 해보아라.

1. 조사목적 : 한신대학교 4학년의 희망 취업 분야 조사

2. 표본설계 : 목표모집단 – 한신대학교 4학년
   조사모집단 – 2019년 1학기 기준 한신대학교 4학년으로 재학 중인 자
   목표표본 – 각 희망 취업 분야에 대한 모비율 추정을 위한 표본 크기 설정을 위해목표 추정오차를 0.05, 신뢰구간을 95%, 모분산이 최대가 되도록 $n_0 = 1.96^2 \frac{PQ}{B^2} = 1.96^2 \frac{0.5 \times 0.5}{0.05^2} = 384.16$ 을 구하고, 모집단 5000/4명(4개 학년이 균등하게 존재한다. 가정)을 가정하여 최종 표본 크기를 구하면 $n = \frac{n_0}{1+n_0/N} = \frac{384.16}{1+384.16/1250} = 293.85$ 의 목표 표본을 설정한다.
   추출방법 – 1차로 학과를 뽑고, 2차로 학생을 뽑는 2단계 층화추출법
   26개 학과 중 40%인 10개 학과를 SRS로 추출하고, 뽑힌 10개 학과에서 추출틀을 이용하여 30명 씩 SRS로 추출
   추출틀 – 각 학과별 4학년 연명부

3. 측정설계 : 성별, 주소, 나이, 학과 등의 인적사항과 조사 목적인 희망 취업 분야를 묻기 위한 설문 문항들을 작성한다.

4. 조사방법 : 표본조사를 위해 면접원이 직접 면접조사를 실시

5. 조사기간 : 2019년 6월 1일 ~ 2019년 6월 30일 기간 동안 실시

#18 표본 크기를 결정하는 방법, 기준에 대해 설명하시오.

표본 크기를 결정할 때는 기본적으로 조서목적과 여건을 고려하며, 일반적으로 표본추출방법, 목표 오차, 신뢰수준, 분포, 비용, 응답률 등을 고려하여 결정한다. 표본 조사에서 예상하는 표본오차에 대한 목표 추정오차(B)를 미리 설정한다. 추정오차 = $|\Theta - \widehat{\Theta}| < B$, 여기서 Θ 는 모수, $\widehat{\Theta}$ 는 표본 추정량을 의미한다. 그 다음 신뢰수준$(1-\alpha)$를 결정한다. $\Pr(|\Theta - \widehat{\Theta}| < B) = 1-\alpha$ 이다. 추정오차와 신뢰구간이 결정되면 목표 추정 오차를 얻을 수 있는 적절한 표본추출방법을 선택하여 표본 크기를 결정한다. 일반적으로 비용보다는 최소 분산을 갖는 조건에서 오차 한계를 달성할 수 있는 표본 크기를 계산한다.

예를 들어 단순확률추출법을 가정할 때 신뢰수준 95%에서 모평균 추정을 위한 표본크기를 결정한다면 복원 추출일 경우 $n_0 = 1.96^2 \frac{\sigma^2}{B^2}$, $(B = 1.96\sigma_{\bar{X}} = 1.96\frac{\sigma}{\sqrt{n}})$ 이 되고, 비복원 추출 혹은 유한모집단의 경우 $n = \frac{n_0}{1+n_0/N}$ 이다. σ 는 과거 조사를 이용하거나 간단한 조사를 통해 구한다.

그러나 현실에선 절대오차를 사용하기 어려워 상대표준오차를 이용하여 표본 크기를 결정한다. $n_0 = \left(\frac{CV_{\hat{\theta}}}{d}\right)^2$ -〉 $n = \frac{n_0}{1+n_0/N}$ 이다. 여기서 CV(표준편차/평균)는 모집단의 변동계수, d는 목표 상대표준오차(추정량의 변이계수), N은 모집단의 크기이다.

#19 현재 조별로 분석하고 있는 내용을 5줄로 요약하여 설명하시오.

#20 스스로 문제를 내고 답하시오.